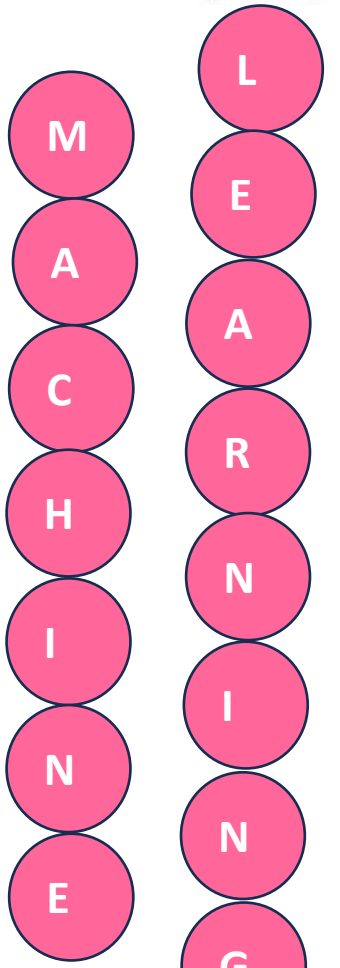
BREAST CANCER CLASSIFICATION

A study using KNN & Logistic Regression Models

Data: www.kaggle.com/datasets/erdemtaha/cancer-data
Colab Link: https://colab.research.google.com/CW2_ML

Name: Sandra Binu
Student ID:22029960
Subject ID: 7PAM2021-0105-2023

Machine Learning in Cancer Diagnosis



aids in early cancer detection by analyzing medical data to identify patterns indicative of malignancy

Algorithms analyze a variety of patient data, such as genetic markers and imaging results, to provide accurate cancer diagnoses.

Classification maximise efficacy by customizing treatment plans based on the unique characteristics of each patient.

accurately detects benign and malignant tumors, reducing the need for needless treatments.

improves patient outcomes by providing individualised care and prompt diagnosis

ML-based diagnosis speeds up the procedure, allowing for quicker and better-informed choices to be made for patient care.

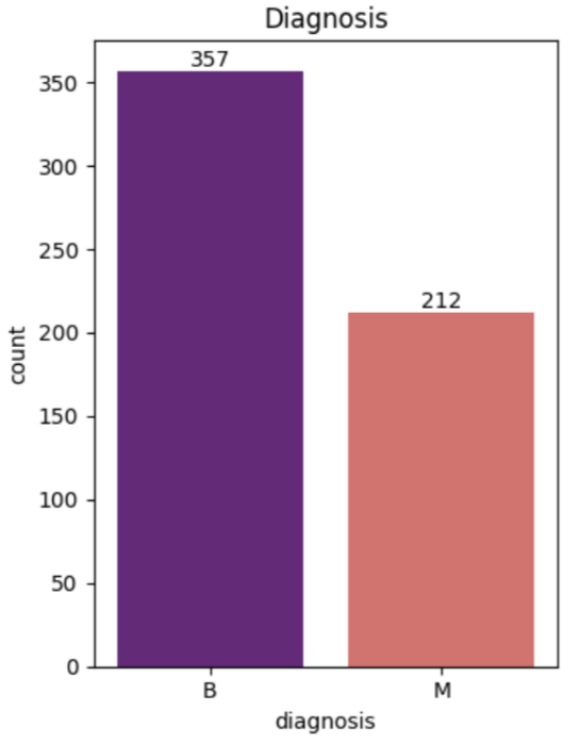
BREAST CANCER DATASET

Total 569 Biopsies

1)radius (mean of distances from center to points on the perimeter)
2)Texture (standard deviation of gray-scale values)
3) Perimeter
4) Area
5) Smoothness (local variation in radius lengths)
6) compactness (perimeter^2 / area - 1.0)
7) concavity (severity of concave portions of the contour)
8) concave points (number of concave portions of the contour)
9) symmetry
10)fractal dimension("coastline approximation“-1)

×

- Mean
- Standard Error
- Worst (mean of the three largest values)



Target Variable: Diagnosis [Benign-357(non-cancerous) or Maliignant-212 (cancerous)]

- Unnamed:32

+

- ID
- diagnosis

+

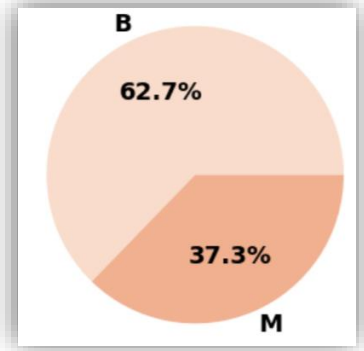
30 Features

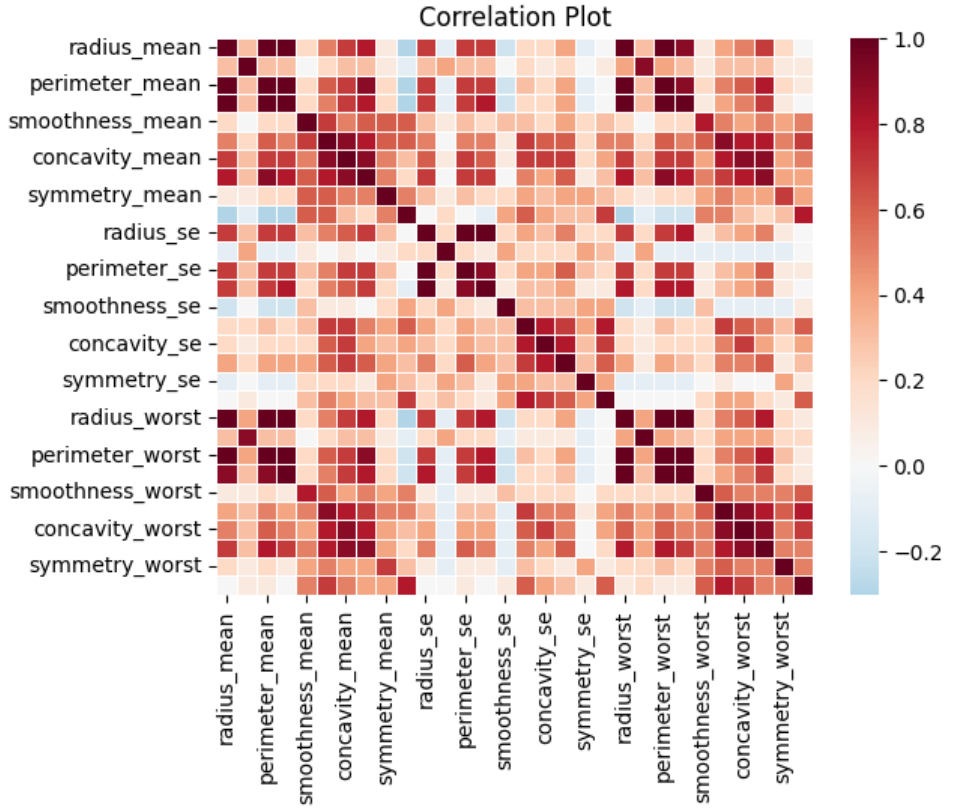
=

33 Columns

Exploratory Data Analysis

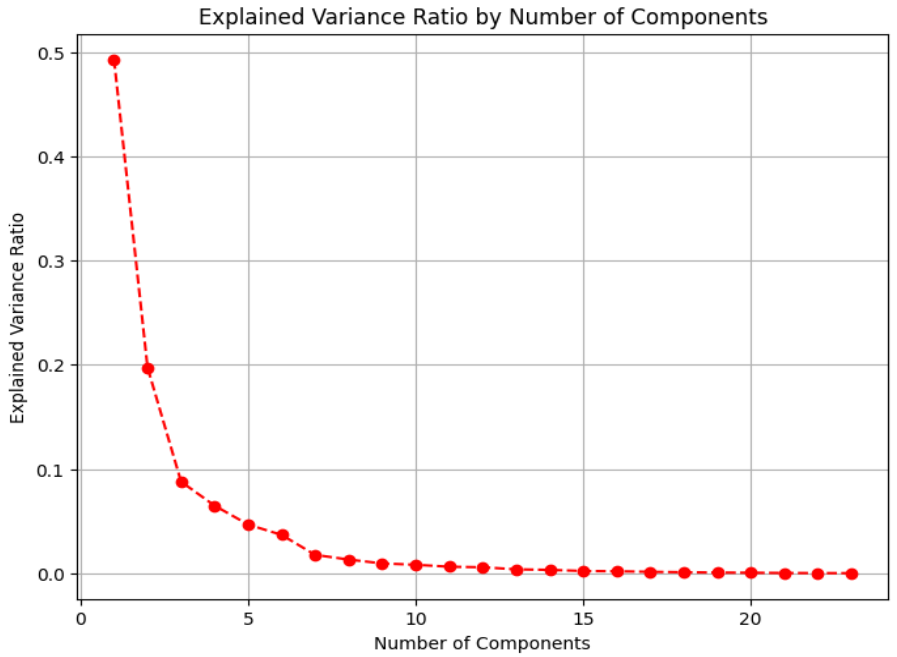
- Dtype of each column checked : ID (int64) , diagnosis(object), others(float64)
- null value count in each column checked: 0 null values except one ‘Unwanted 32:’ column.
- data.describe() explains the count and other relevant details of other features(569)
- count of each class in the target variable (Benign(357) and Malignant(212)) & pieplot gives its distribution.
- distribution of all the features by diagnosis (target variable)
- Multicollinearity and Correlation of all features checked using sns.pairplot() and sns.heatmap()





Data Preprocessing

- Drop the unwanted columns(id,unnamed:32) from the data and set separate variables for target and features.
- StandardScaler() to standardize all the features into same scale and LabelEncoder() to encode the target variable(B-0 & M-1)
- Since high multicollinearity and correlation, apply Dimensionality Reduction using PCA for the features.
  - ❖Find optimum number of Principle Components using Scree Plot & transform scaled features data.
- Split the data as train(455 instances) & test(114 instances)



Number of components taken is 14 as it covers almost 80% variance of the data and contributes to model efficiency

K- Nearest Neighbors

Classification & Regression

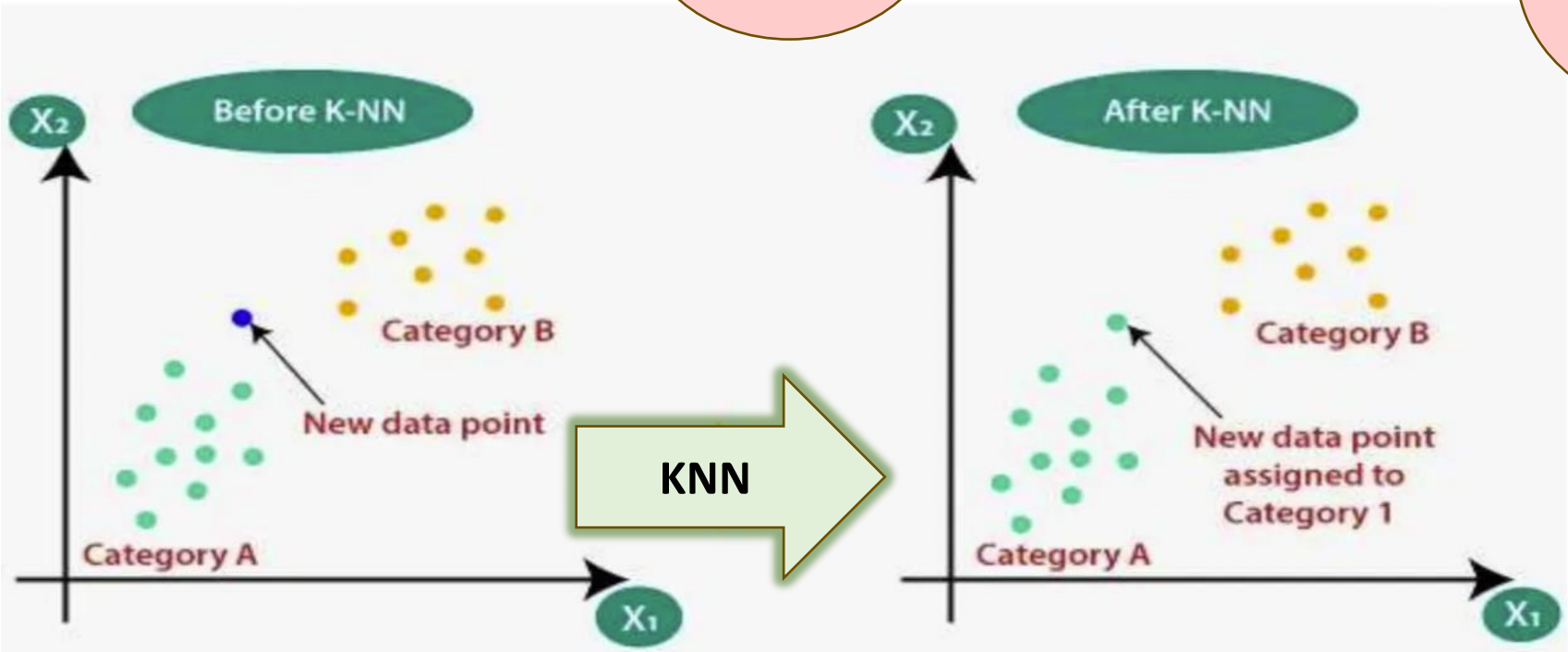
Non-parametric

- Classifies a new data point based on the majority class of its k nearest neighbors.
- Uses distance metrics like Euclidean distance, Manhattan

Supervised Learning

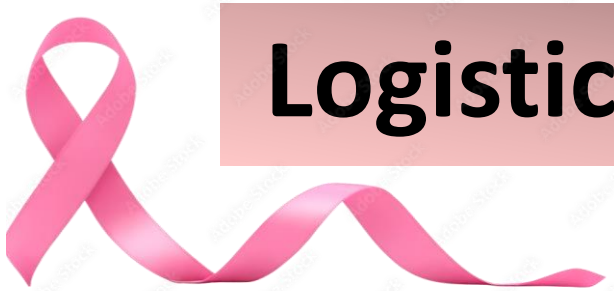
Lazy Learning

Instance-based Learning



Logistic Regression

Linear model for Binary Classification

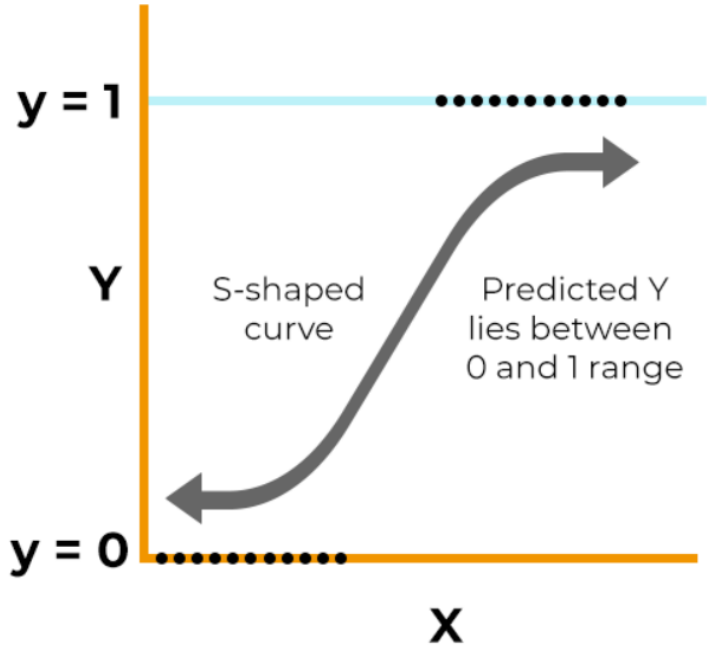


- The probability of dependent variables is modeled as a function of independent variable

- maps the linear combination of input features to a probability between 0 and 1 using the logistic function (sigmoid function)

outputs probabilities that a given input belongs to a specific class

assumes a linear relationship between the log-odds of the result and the input features



$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}$$

x : input value

y :predicted output

b_0 : bias or intercept term

b_1 : coefficient of x

Hyperparameter Tuning using GridSearchCV

Hyperparameter Tuning : Optimize the performance of the model by selecting the best hyperparameters.
GridSearchCV : Utilized for exhaustive search over designated parameter grid & cross-validation to prevent overfitting.

KNN

Best Parameters:
'metric': '**manhattan**',
'n_neighbors': **5**,
'weights': '**uniform**'

Parameter Grid:
- **n_neighbors**: Number of neighbors to consider.
- **metric**: Distance metric to measure similarity (Euclidean or Manhattan).
- **weights**: Scheme for weighting neighbors (uniform or distance)

➢Test accuracy before Tuning: **0.9474**
➢Test accuracy after Tuning: **0.9561**

Logistic Regression

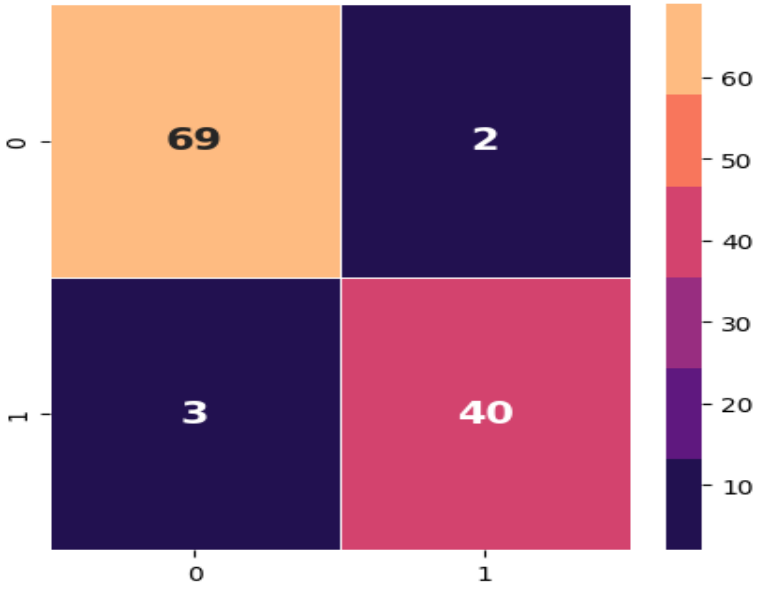
Best Parameters:
'C': **0.1**,
'penalty': '**l2**',
'solver': '**liblinear**'

Parameter Grid:
- **C**: Regularization strength.
- **penalty**: Type of penalty (L1 or L2).
- **solver**: Algorithm to use in the optimization problem

➢Test accuracy before Tuning: **0.9912**
➢Test accuracy after Tuning: **0.9912**

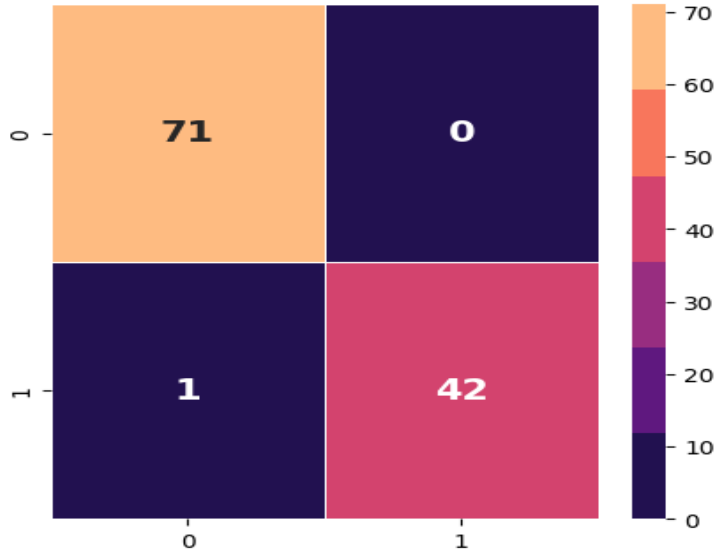
PERFORMANCE OF KNN AND LOGISTIC REGRESSION ON THE DATA

KNN Confusion Matrix



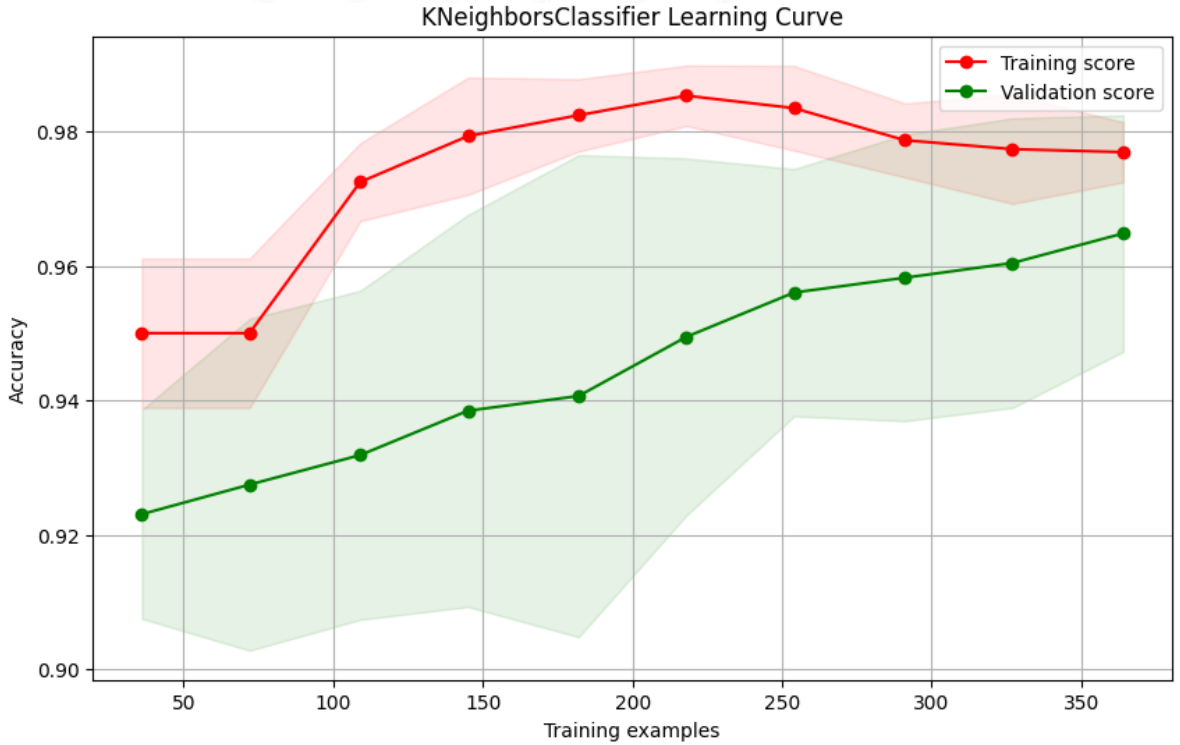
- When compared to KNN, the LR model has higher accuracy and reliability in classifying malignant cases, with only one false negative and no false positives.
- Precision and recall are critical in cancer diagnosis.
- While LR's higher recall guarantees early detection and treatment, its higher precision eliminates unnecessary treatments.

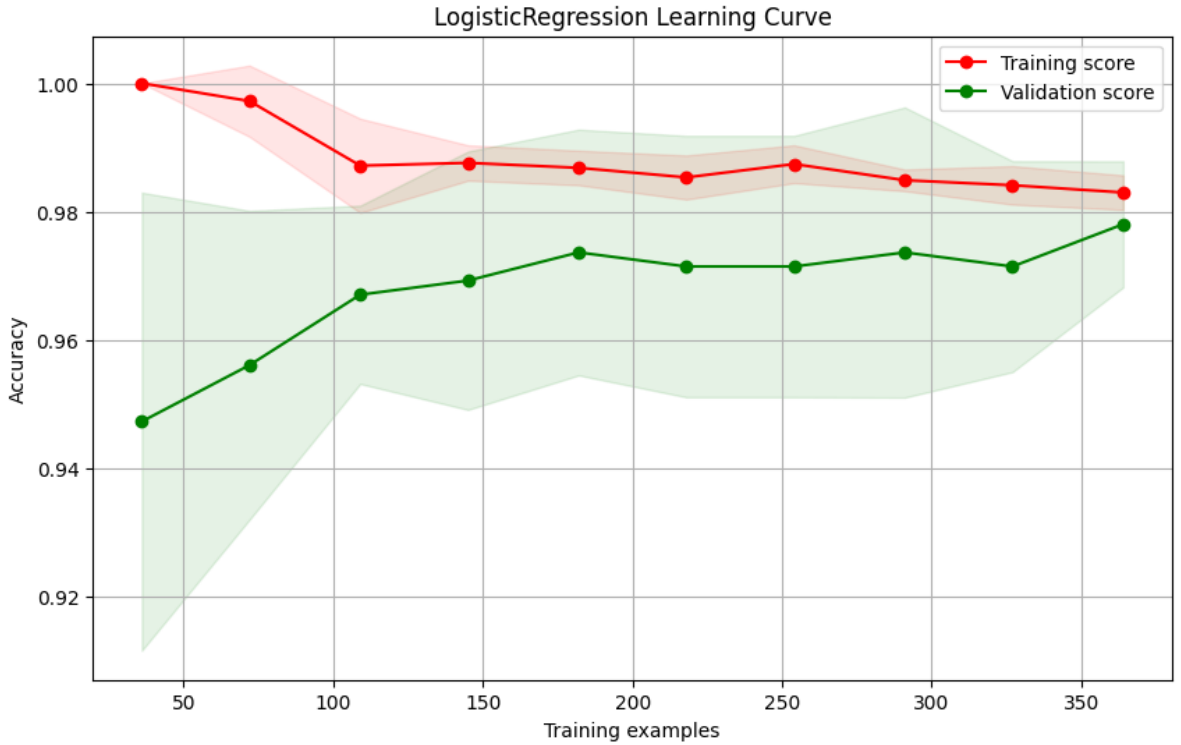
Logistic Regression Confusion Matrix



	KNN				Logistic Regression			
	Precision	**Recall**	**F1-score**	**Support**	**Precision**	**Recall**	**F1-score**	**Support**
Benign	0.96	0.97	0.97	71	0.99	1.00	0.99	71
Malignant	0.95	0.93	0.94	43	1.00	0.98	0.99	43

Performance of Models on Data





- Strong performance on the training data is indicated by a rapid increase in the training score.
- A higher training score indicates greater training accuracy but lower accuracy on unseen data compared to LR than validation score, which may be a sign of overfitting and limited generalization to new data.
- Steady increase in training score indicates effective learning from training data.
- Plateauing validation score suggests potential overfitting and challenges in generalization to unseen data.

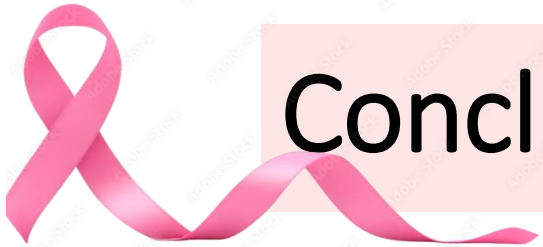
Limitations

Bias in data collection or model training can lead to biased predictions, potentially exacerbating existing disparities in healthcare especially a case like cancer diagnosis.

ML models may overfit to training data, capturing noise or outliers and reducing their ability to generalize data. That might be the reason why LR model is not showing further improvement after hyperparameter tuning.

Other ML models like SVM, Random Forest and Neural Networks may be studies to give better comparison than KNN.

Rigorous validation and regulatory approval processes are necessary to ensure the safety and efficacy of ML-based diagnostic tools before clinical implementation.



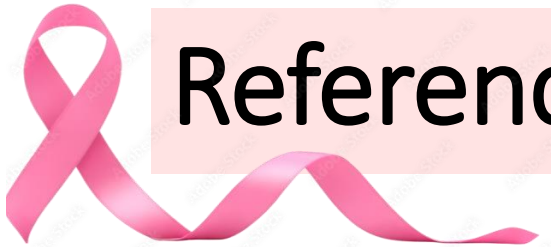
Conclusions

Study identifies Logistic Regression to be most successful in breast cancer classification.
(99.21% accuracy rate, and the least number of fault predictions)

KNN lacks interpretability, making it less useful for understanding cancer classification factors.

LR's superior precision, recall, and interpretability make it the preferred choice for cancer prediction.

LR’s accuracy(0.99) and ability to minimize false positives/negatives are crucial for accurate diagnosis and optimal patient care.



References

- https://bmcmedresmethodol.biomedcentral.com/articles/
- https://medium.com/@hammad.ai/3-ways-to-detect-multicollinearity-in-your-dataset
- https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063
- https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- https://www.geeksforgeeks.org/logistic-regression-vs-k-nearest-neighbors-in-machine-learning/
- https://www.jeremyjordan.me/hyperparameter-tuning/
- https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/
- https://datatron.com/understanding-the-confusion-matrix-for-model-evaluation-monitoring